Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 179-182 recibido 15-03-2019 revisado 09-05-2019 aceptado 14-05-2019

Social-Univ 2.0: Tecnologías del Lenguaje Humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante

Social-Univ 2.0: Human Languages Tecnologies applied to the omnichannel monitoring of the University of Alicante's digital social environment

Isabel Moreno[1,2], Javi Fernández[2], Yoan Gutiérrez[1]
[1]Instituto Universitario de Investigación Informática, Universidad de Alicante
[2]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apdo. de Correos 99 E-03080, Alicante, Spain
{imoreno, javifm, ygutierrez}@dlsi.ua.es

Resumen: Este proyecto viene motivado por la necesidad de conocer el ecosistema digital social universitario de aquellas universidades con presencia en redes sociales, y más concretamente, de la Universidad de Alicante. Para ello, las Tecnologías del Lenguaje Humano juegan un papel fundamental, ya que se utilizan para extraer meta-datos sobre comentarios de las redes sociales y representar tanto perfiles de actores sociales (empresas, instituciones, personas, etc.) como la relación social que existe entre ellos en un determinado período de tiempo. El resultado de este proyecto es facilitar el estudio de fenómenos como: las interacciones sociales y sus matices; analizar grupos sociales y sus comportamientos (i.e. patrones de su crecimiento y decrecimiento); identificación de actores sociales relevantes e influyentes; identificar temas comunes en determinados grupos sociales; y la medición de campañas y eventos a través de los medios sociales.
Palabras clave: Tecnologías del Lenguaje Humano, Universidad, Social, Relaciones

Abstract: This project is motivated by the need to know the digital social university ecosystem of those universities present in social networks. More specifically, the focus is the University of Alicante. To that end, Human Language Technologies (TLH) play a fundamental role. TLH are employed not only to extract meta-data from comments from social networks, but also to represent profiles of social actors (companies, institutions, people, etc.) as well as the social relation that exists between them in a certain period of time. The outcome of this project is to ease the study of phenomena such as: social interactions and its nuances; analyse social groups and their behaviours (i.e. their growth and decline patterns); identification of relevant and influential social actors; identify common themes in certain social groups; and measurement of campaigns and events through social media.
Keywords: Human Language Tecnologies, University, Social, Relations

1 Introducción

Actualmente, Internet cuenta con más de 4000 millones de usuarios conectados[1]. La Web 2.0 (o Web social) es uno de los escenarios de Internet donde los usuarios juegan un papel primordial, pues ya no sólo consumen información, si no que los usuarios también forman parte activa en la creación de contenidos. Estos pueden participar, interactuar e intercambiar información con otros usuarios en redes sociales. Por ejemplo si se analizan los datos de Twitter, una de las redes sociales más conocidas, observamos que cuenta con más de 344 millones de usuarios activos[2], que generan 500 millones de tweets al día[3].

Dicha información textual se encuentra en diversas fuentes de información de distinta naturaleza y en distintos idiomas. Estos fac-

[1] `https://www.internetworldstats.com/stats.htm` (Febrero 2019)

[2] `http://www.internetlivestats.com/watch/twitter-users/` (Febrero 2019)

[3] `http://www.internetlivestats.com/twitter-statistics/` (Febrero 2019)

ISSN 1135-5948. DOI 10.26342/2019-63-25

tores, unidos a la redundancia y a la información contradictoria, suponen que los usuarios inviertan mucho más tiempo de lo deseado analizando la información para seleccionar aquella que sea de su interés.

Por lo tanto, es necesario reducir el tiempo que los usuarios invierten en analizar grandes cantidades de información formuladas en lenguaje natural. Esto se consigue gracias a las Tecnologías del Lenguaje Humano (TLH) que se encargan de procesar automáticamente el lenguaje humano de manera eficiente y efectiva.

Para su correcto funcionamiento se ha mejorado la búsqueda, recuperación y extracción de información (Irfan et al., 2015; Dalvi, Cohen, y Callan, 2012; El-Helw, Farid, y Ilyas, 2012) o la detección y minería de opiniones (Gutiérrez, Tomás, y Fernández, 2015; Ravi y Ravi, 2015; Montoyo, Martínez-Barco, y Balahur, 2012; Mihalcea, Banea, y Wiebe, 2012; Thelwall y Buckley, 2013), junto con los procesos intermedios involucrados en cada una, como el análisis semántico (Li y Joshi, 2012; Gutiérrez, Vázquez, y Montoyo, 2017).

A pesar de esa mejora, es evidente la necesidad de integrar todos los procesos de TLH anteriores en una sola plataforma. De esta forma, la información de redes sociales que necesita el usuario, será recuperada y procesada para presentársela de manera adecuada y flexible para su posterior análisis.

2 Estado del arte

En la actualidad existen infraestructuras analíticas que de alguna forma incluyen TLH. Algunos ejemplos serían: (i) Atribus[4] rastrea, busca, recoge, filtra y devuelve todo lo que se está diciendo de un cliente en la red a partir de las palabras clave en tiempo real; (ii) Natural Opinions[5] analiza todo lo que se está diciendo en cada momento en Internet sobre una persona, una marca, una institución o un producto, y detecta automáticamente las entidades, conceptos y opiniones más relevantes; (iii) Textalytics[6] extrae elementos con significado de cualquier contenido y lo estructura para que procesarlo y gestionarlo fácilmente; (iv) Sentiment viz[7] estima y visualiza el sentimiento asociado a textos cortos e incompletos; (v) Tweet Reach[8] permite obtener informes estadísticos a partir de analizar twitter; (vii) SocialBro[9] gestiona y analiza comunidades de Twitter para que los profesionales del Marketing analicen a fondo sus contactos y definan sus estrategias consecuentemente; y (viii) SumAll[10] obtiene estadísticas sobre seguidores de varias redes sociales como Facebook o Instagram y genera gráficas de intervalos de tiempo (días, semanas, meses) con números de me gustas, mensajes, localidades, etc.

Sin embargo, si nos enfocamos en el uso de las TLH a día de hoy, no encontramos infraestructuras analíticas que analicen el impacto que las universidades tienen en las redes sociales. Por tanto, si deseáramos conocer el ecosistema digital social en el entorno de una universidad, sería necesario emplear cuantiosos recursos, humanos y temporales, en entender cada noticia y comentario que se emite sobre esa universidad. Después, podríamos crear perfiles e identificar roles de todos aquellos actores y eventos que pudieran incidir en su reputación a nivel social y global.

3 Propuesta

Ante la falta de tecnologías que resuelvan esta problemática, el Grupo de Procesamiento del Lenguaje Natural y Sistemas de Información (GPLSI) se plantea este proyecto, donde se persigue como objetivo estudiar, a través del uso de las TLHs, el entorno digital-social universitario. Por este motivo, se elige como caso de estudio la Universidad de Alicante (UA) y entidades relacionadas con esta institución. El objetivo principal se centra en el uso de las TLHs para extraer meta-datos sobre comentarios de las redes sociales y representar tanto perfiles de actores sociales (empresas instituciones, personas, etc.) como la relación social que existe entre ellos en un determinado período de tiempo. La Figura 1 nos proporciona una idea gráfica del resultado esperado.

Para alcanzar nuestras metas requerimos del procesamiento de grandes cantidades de información proveniente de las redes sociales, por lo que se hace uso de plataformas ya existentes en el GPLSI. Entre ellas, destacamos la plataforma Social Analytics (Fernández et al., 2017), con su ayuda podemos obtener

[4] http://www.atribus.com (Febrero 2019)

[5] http://www.bitext.com (Febrero 2019)

[6] https://www.meaningcloud.com (Febrero 2019)

[7] http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app (Febrero 2019)

[8] http://tweetreach.com (Febrero 2019)

[9] http://es.socialbro.com (Febrero 2019)

[10] https://sumall.com (Febrero 2019)

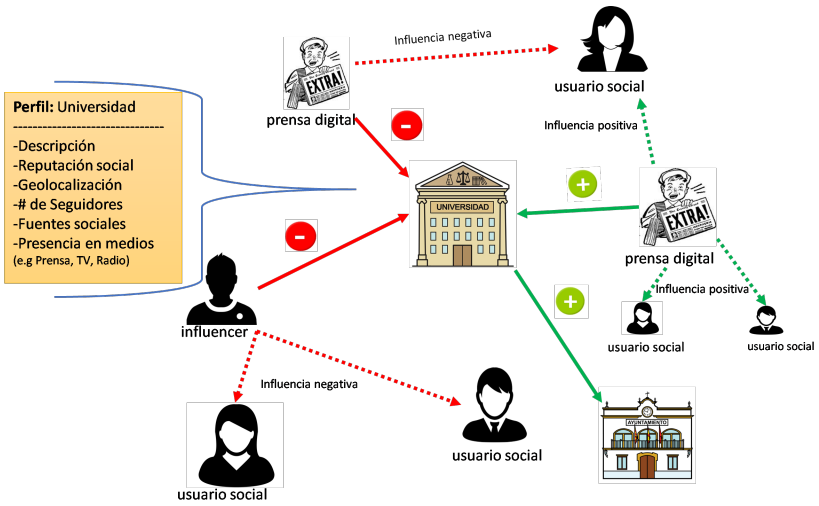

Figura 1: Perfiles y relaciones sociales en la Web 2.0

contenido de la Web Social y extraer metadatos sobre cualquier entidad digital, en este caso de estudio la UA, desde un punto de vista analítico, tras un intenso procesamiento de mensajes de usuarios sociales. A continuación, el análisis de sentimientos y emociones intervienen como las claves para valorar las diferentes posturas (valoración, opinión, etc.) que cada actor tiene respecto a otros. Finalmente se diseñan y aplican algoritmos que permiten obtener modelos de representación tanto de perfiles de actores sociales (empresas, instituciones, personas, etc.) como la relación social que existe entre ellos en un determinado período de tiempo.

Para que el proyecto se lleve a cabo con éxito, son precisas las siguientes tareas:

1. Análisis y definición de tecnologías: Esta tarea tiene como objetivo analizar los requerimientos de este proyecto, estudiando en profundidad los tipos de aplicaciones, recursos, herramientas y tecnologías necesarias, así como sus disponibilidades y viabilidades en el desarrollo de ellas. Se definen también las guías y formato a seguir en el desarrollo de TLH necesarias. Además, se establece el modo de visualización para presentar los datos.

2. Recuperación y extracción de información: En esta tarea se reutilizan y desarrollan TLHs que permitan rastrear, recuperar y extraer información de distintas fuentes (webs, foros, blogs, redes sociales, etc.), formatos (JSON, XML, HTML, etc.) e idiomas (inicialmente inglés y español).

3. Minería y procesamiento de textos: En esta tarea reutilizan y desarrollan las TLHs necesarias para que, a partir de textos escritos en lenguaje humano, estos sean analizados y posteriormente clasificados y estructurados para ser almacenados considerando tecnologías de Minería de Datos. La clasificación resultante es interpretada según: el significado de la unidad de información; los dominios al que se refiere el contexto; la determinación si es o no una opinión (información objetiva versus subjetiva); polaridad emocional implicada; términos relevantes detectados y contabilizados; la geolocalización de contenido generado por usuarios y de autores; etc. Además, dentro de esta tarea se tienen en cuenta qué aspectos pueden influir en la calidad del contenido de la información, como por ejemplo, información redundante, contradictoria, inconsistente, etc., y que pueden afectar al texto generado.

4. Presentación del contenido resultante: En esta tarea se analizan, reutilizan y desarrollan enfoques que permitan representar el contenido procesado automáticamente de un modo útil y sencillo para ser interpretados por usuarios finales. Es decir, representar tanto perfiles de actores sociales como la relación social que existe entre los actores en un determinado período de tiempo.

5. Integración de tecnologías: En esta tarea los mejores enfoques de investigación analizados en las tareas previas se inte-

gran en un único proceso. Esto permite tener como entrada unas necesidades de información, y generar como salida, un paquete con información anotada a distintos niveles, que será la que dé lugar al contenido finalmente mostrado a usuarios finales.

6. Evaluación de los resultados: La evaluación es un aspecto clave y crucial para el proyecto. Se realizan dos tipos de evaluaciones: (i) una evaluación intrínseca de cada una de los enfoques analizados y propuestos para cada tarea involucrada en el proyecto en base a sistemas de referencia, y (ii) una evaluación extrínseca para medir la utilidad de las tecnologías desarrolladas como un todo. Para ambas se utilizan métodos cuantitativos, basados en las medidas tradicionales de precisión y cobertura, y cualitativos, que permitan analizar criterios para los que no existen todavía medidas automáticas (por ejemplo, coherencia de los resultados mostrados).

4 Resultados

El resultado tecnológico de este proyecto es una herramienta llamada *Social-Univ 2.0* que involucra las tecnologías antes mencionadas. En un futuro, dicha herramienta podría adaptarse a aplicaciones móviles y el UACloud[11] de la UA. *Social-Univ 2.0* requerirá para su funcionamiento del uso de tecnologías de minería de textos sociales provenientes del sistema Social Analytics.

Con las tecnologías resultantes de este proyecto es posible estudiar el entorno social de las entidades sociales digitales, en concreto la universidad de Alicante. Esto incluye: estudiar las interacciones sociales y sus matices; analizar grupos sociales y sus comportamientos (i.e. crecimiento, decrecimiento); identificación de actores sociales relevantes e influyentes; identificar temas comunes en determinados grupos sociales; y la medición de campañas y eventos a través de medios sociales.

Agradecimientos

Este proyecto con referencia ENCARGOINTERNO5-19EN: "Social-Univ 2.0: Tecnologías del Lenguaje Humano, aplicación para la monitorización omnicanal del entorno social de la Universidad de Alicante", está financiado por la Universidad de Alicante.

[11] `https://cvnet.cpd.ua.es/uacloud/` (Febrero 2019)

Bibliografía

Dalvi, B. B., W. W. Cohen, y J. P. Callan. 2012. Websets: extracting sets of entities from the web using unsupervised information extraction. En *WSDM*.

El-Helw, A., M. H. Farid, y I. F. Ilyas. 2012. Just-in-time information extraction using extraction views. En *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, páginas 613–616, New York, NY, USA. ACM.

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y Á. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58:141–148.

Gutiérrez, Y., D. Tomás, y J. Fernández. 2015. Benefits of using ranking skip-gram techniques for opinion mining approaches. En *eChallenges e-2015 Conference*, páginas 1–10. IEEE.

Gutiérrez, Y., S. Vázquez, y A. Montoyo. 2017. Spreading semantic information by word sense disambiguation. *Knowledge-Based Systems*, 132:47 – 61.

Irfan, R., C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, y et al. 2015. A survey on text mining in social networks. *The Knowledge Engineering Review*, 30(2):157–170.

Li, Y. y K. D. Joshi. 2012. The state of social computing research: A literature review and synthesis using the latent semantic analysis approach. En *AMCIS*.

Mihalcea, R., C. Banea, y J. Wiebe. 2012. Multilingual subjectivity and sentiment analysis. En *Tutorial Abstracts of ACL 2012*, ACL '12, páginas 4–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Montoyo, A., P. Martínez-Barco, y A. Balahur. 2012. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675 – 679.

Ravi, K. y V. Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14 – 46.

Thelwall, M. y K. Buckley. 2013. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617.